arXiv:cond-mat/0202090v1 [cond-mat.stat-mech] 6 Feb 2002

Optimal prediction of folding rates and transition state placement from native state geometry

Cristian Micheletti
International School for Advanced Studies (S.I.S.S.A.) and INFM, Via Beirut 2-4, 34014 Trieste, Italy
(Dated: February 1, 2008)

A variety of experimental and theoretical studies have established that the folding process of monomeric proteins is strongly influenced by the topology of the native state. In particular, folding times have been shown to correlate well with the contact order, a measure of contact locality. Our investigation focuses on identifying additional topologic properties that correlate with experimentally measurable quantities, such as folding rates and transition state placement, for both two- and three-state folders. The validation against data from forty experiments shows that a particular topologic property which measures the interdepedence of contacts, termed cliquishness or clustering coefficient, can account with significant accuracy both for the transition state placement and especially for folding rates, the linear correlation coefficient being $r = 0.71$. This result can be further improved to $r = 0.74$, by optimally combining the distinct topologic information captured by cliquishness and contact order.

I. INTRODUCTION

In the past three decades, there has been a growing effort of the scientific community for studying and understanding the principles that govern the folding process of a sequence of amino acids in the corresponding native structure[1,2,3]. In recent years, several proteins, in particular those folding via a two-state mechanism[4] have provided an extraordinary benchmark for experimental and theoretical characterization of the folding pathways. The significant amount of experimental data available for several structurally unrelated proteins[4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24], has opened the possibility to identify and isolate the factors that influence the folding rate. Besides considering detailed chemical interaction, such as those affecting free-energy barriers, an appealing and elegant line of investigation has focused on the effects of the native state structure on the folding process[25,26,27,28].

From a qualitative point of view, the influences of structural effects was traditionally summarised in the tenet that proteins with high helical content fold faster than proteins with mixed alpha/beta content, the slowest folding being for the all-beta ones. This useful and intuitive rule of thumb, fails to account for the very different rates observed between proteins in each of the alpha, alpha/beta or beta families[5,29,30,31]. A deep insight into this problem was provided by the work of Plaxco *et al.*[25] who introduced the concept of contact order, which captures, quantitatively, features beyond the mere secondary structure motifs. The highly significant correlation of contact order and experimental folding rates shows the extent to which the mere topology of native state can influence the folding process. However, the highly organised native structure of proteins is too rich to be captured by a single parameter such as the contact order. Indeed, the latter cannot account in the same satisfactory way for the transition state placement, three-state folding rates or the diversity of folding rates among structurally similar proteins[32].

In the present study we investigate how the topology of the native state can be further exploited to provide optimised predictions for protein folding rates and the transition state placement. To do so we consider, among others, one particular topological descriptor that is crucial for characterising the connection and interactions of native contacts: the clustering coefficient , or cliquishness. Such parameter, heavily studied in the context of graph theory[33,34,35] is shown to have highly significant correlation with folding rates. The advantage of using this topologic descriptor is that it allows to capture the cooperative formation of native interactions, as proved by its statistically relevant correlations with the transition state placement. Further, we discuss how the different topologic aspects captured by the cliquishness and contact order can be combined to yield optimal correlations higher than for the individual descriptors.

II. THEORY AND RESULTS

Customarily, at the heart of theoretical or numerical studies of topology-based folding models is the contact matrix (or map)[36] which will be used extensively also in the present context. The generic entry of the contact map, Δ_{ij}, takes on the value 1 if residues i and j are in contact and zero otherwise. Several criteria can be adopted to define a contact; in the present study we consider two amino acids in interaction if any pair of heavy atoms in the two amino acids are at a distance below a certain cutoff, d. All values of d between 3.5 Å and 8 Å have been considered and reported. The contact map provides a representation for the spatial distribution of contacts in the native structures

that is both concise and often reversible (since native structures can be recovered when appropriate values of d are used). Plaxco and coworkers[25] have used the contact map to describe and characterize the presence and organization of secondary motifs in protein structures. The parameter that was introduced, the relative contact order, provides a measure of the average sequence separation of contacting residues and is defined as

$$\text{relative contact order} = \frac{1}{L}\frac{\sum_{i\neq j}\Delta_{ij}\ |i-j|\ w_{ij}}{\sum_{i\neq j}\Delta_{ij}\ w_{ij}}, \tag{1}$$

where i and j run over the sequence indeces, w_{ij} is the contact degeneracy (i.e. the number of pairs of heavy atoms in interaction) and L is the protein length. Remarkably, the contact order was shown have a highly significant linear correlation with experimental folding rates. The result of Plaxco and coworkers can be explained, *a posteriori*, with intuitive arguments: a high contact order corresponds to few local interactions. One may thus expect that the route from the unfolded ensemble to the native state is slow, being hindered by the overcoming of several barriers[37,38,39,40,41] due to spacial restraints, as recently analysed by Debe and Goddard[26] and previously by Chan and Dill[42] and also observed in topology-based numeric studies[43]. These considerations are based purely on geometric arguments and do not take into account the influence of specific interactions between the residues. In principle, the latter may well override the topological influence on the folding process, but surprisingly, as remarked in a recent review article[44] this is often not the case[43,45,46,47,48,49,50].

Our aim is to exploit as much as possible the topologic information contained in the native state to improve both the accuracy of predictions for folding rates and gain more fundamental insight into the process. To this purpose we have considered additional topologic descriptors besides the contact order. The one that appeared most significant is a parameter termed cliquishness or clustering coefficient[33,34,35]. For a given site, i, the cliquishness is defined as:

$$\text{cliquishness}(i) = \frac{\sum_{j\neq l}\Delta_{ij}\Delta_{il}\Delta_{lj}}{\sum_{j\neq l}\Delta_{ij}\Delta_{il}} = \frac{\sum_{j\neq l}\Delta_{ij}\Delta_{il}\Delta_{lj}}{N_c(N_c-1)/2}, \tag{2}$$

where N_c is the number of contacts to which site i takes part to. As for the contact order, also the cliquishness has an intuitive meaning; in fact it provides a measure of the extent to which different sites interacting with i are also interacting with each other. Of course, the cliquishness is properly defined only if site i is connected to, at least, two other sites. To ensure this, we included also the covalently bonded interactions $[i, i\pm1]$ in (2). The importance of taking the cliquishness into account for discriminating fast/slow folders can be anticipated since a higher interdependency of contacts (large cliquishness) will likely result in a more cooperative folding process. In fact, the formation of a fraction of interactions will result in the establishment of a whole network of them. Consistently with this intuitive picture one should also expect that a large/small cliquishness will affect in different ways the amount of native-like content of the transition state.

We have tested and verified these expectations by calculating the average cliquishness for 40 proteins for which folding rates and transition state placement, θ_m, have been measured. θ_m is deduced from the variation of folding/refolding rates upon change of denaturant concentration ($m_F^\dagger$ and $m_U^\dagger$) and provides an indirect indication of how much the solvent-exposed surface of the transition state is similar to that of the native one. θ_m ranges between 0 and 1; higher values denote stronger similarity with the native state. It is worth pointing out that, although the model underlying the calculation of θ_m relies on a two-state analysis, an effective θ_m can be inferred for three-state folders as well[5]. Since reliable θ_m's are not available for all proteins, the number of entries used to correlate the cliquishness and θ_m (see Tables I and II) is slightly smaller than that used for tge logarigthm of refolding rates, $\ln K_F$. The set of proteins used, shown in Tables I and II, was built up from experimental data collected in previous studies and predictions (often topology-based) of folding rates[5,25,26,27,28].

As indicated, the entries include both two-state and three-state folders, proteins belonging to the same structural family as well as proteins under different experimental conditions. This allows to examine to what extent predicted folding rates are consistent with the wide variations of folding velocities observed in structurally-related proteins and in different experimental conditions. As discussed in detail below, when the comprehensive set of Table I and II is used, the correlation found between cliquishness and folding rates is 0.71, with a statistical significance of $t = 10^{-5}$, more relevant than the one between a suitably defined contact order and folding velocities ($r = 0.66$, $t = 5\cdot10^{-5}$). As will be shown, the predicting power of the two quantities can be combined to achieve the optimal correlation of 0.74.

The prediction of the transition state placement, turns out to be more difficult when either of the two topologic parameters is used. While for the contact order it is equal to 0.23, the cliquishness yields the value of 0.48 which is not significantly improved by combining the two descriptors. Though the linear correlation of the clustering coefficient and the transition state placement is not as high as for the folding-rate case it is nevertheless statistically meaningful, having a probability of 0.004 to have arisen by chance.

A. Two- and three-state folders

Before considering the more general case of all entries in Tables I and II, we focus on two-state folders, i.e. proteins with a cooperative (all-or-none) transition between the unfolded and folded states. The neatness of this process, due to the absence of any significantly populated intermediate state, makes them ideal candidates for identifying and isolating the factors that influence the folding rate.

In the present context this separate test is important since it appears that the relative contact order is a much stronger descriptor for two-state folders, than for the general case. As a matter of fact, when both two- and three-state folders are considered, the influence of the average sequence separation of native contacts on folding properties is better captured by a different version of the contact order, which we shall term "absolute", obtained when the r.h.s. of eq. (1) is not divided by L:

$$\text{absolute contact order} = \frac{\sum_{i \neq j} \Delta_{ij} \ |i-j| \ w_{ij}}{\sum_{i \neq j} \Delta_{ij} \ w_{ij}} \ . \tag{3}$$

In the following we shall report and compare the performance of both parameters; furthermore we shall always consider the absolute value of the linear correlation coefficients, $|r|$, without regard to its sign, which can be easily inferred from the plots.

The original definition of contact order has an unrivaled performance in the prediction of folding rates for the two-state folders of Table I. As visible in Fig. 1, it gives a stable correlation for cutoffs in the range 5 Å$\leq d \leq$ 7 Å, with the maximum value of $r = 0.80$ for the cutoff $d = 4.5$ Å. The statistical significance of such correlation can be quantified through a calculation of the probability, t, to observe by pure chance a correlation higher than the measured one (in modulus). The standard model underlying such estimates relies on the hypothesys of normal distribution of the deviates of the correlated quantities. As a rule of thumb, the upper value of $t = 0.05$ is taken as a threshold for statistically meaningful correlations. For the value of $r = 0.80$ reported above, this probability is $t = 3 \cdot 10^{-5}$, which is, therefore, extremely significant.

Consistently with previous results, we found that the transition state placement is a much more elusive quantity to predict than folding rates. In fact, all topologic descriptors yield a poorer correlation compared to $\ln K_F$ (see Fig. 2). For the relative contact order, the best r is 0.48 (for $d = 6.0$) with an associated $t = 0.02$. As anticipated, the performance of the absolute contact order in this particular context is significantly inferior then the relative one (see Figs. 1 and 2) and hence will not be further commented.

Concerning the performance of the novel parameter under scrutiny, the cliquishness, it can be seen from Figs. 1 and 2 that it is statistically meaningful for both folding rates and transition state placement. There are, however, significant differences with respect to the contact order. For folding rates the optimal r is 0.67 ($d = 4.6$ Å) and the associates value of t is 5 10^{-4}, one order of magnitude larger than for the relative contact order. For θ_m the situation is reversed since the optimal value of $r = 0.58$ (for $d = 3.8$ Å) has the statistical relevance of $t = 0.004$, with a marked improvement over the previous case.

It is also interesting to note that cliquishness-based correlations have a non-trivial dependence on the cutoff d. In fact, due to the overall compactness and steric effects, the degree of dispersion of the cliquishness values for different sites in the same or different proteins is much more limited compared, e.g. to the average sequence separation of contacts. This leads to the observed decay of the correlations when the cutoff d is increased.

The applicability of topology-based models are not limited to two-state folder, but can be extended to include three-state folders as well[45,46]. Despite the addition of the 11 entries corresponding to three-state folders, the performance of cliquishness-based predictions for folding rates and θ_m improves from the values reported for two-state folders. As shown in Figs 3 and 4 the associated optimal correlations for $\ln K_F$ and θ_m are $r = 0.71$ and 0.49, again observed for the same cutoff values (d) mentioned for the two-state case. The corresponding statistical significances are now, $t = 1 \cdot 10^{-5}$ and $t = 0.004$, which, despite the enlargement of the experimental set, show even an improvement over the two-state case.

From Figs. 3 and 4 it can be noticed that the performance of the relative contact order is noticeably poorer than the absolute contact order which, being a much better descriptor, becomes the focus of our analysis. The corresponding measured correlations are, in fact, $r = 0.66$ for $\ln K_F$ and $r = 0.20$ for θ_m with corresponding values of $t = 5 \cdot 10^{-5}$ and $t = 0.23$.

A direct comparison of how the clustering coefficient and the absolute contact order correlate with $\ln K_F$ and θ_m can be made by inspecting the plots of Figs. 5 and 6. It is worth pointing out that the analysis of the deviations from the linear trends of Figs. 5 and 6 reveals that a particular protein, 1urn, is among the top outliers for both cliquishness and contact order-based analysis, although no simple explanations is available for this singular behaviour. Although for both folding parameters the cliquishness gives a more significant correlation than contact order, the difference is

particularly dramatic for the transition state placement which is notoriously difficult to capture with topology-based predictions[25].

An important conclusion stemming out of this observation is that the transition state structure (and hence θ_m) is more influenced by the degree of interdependency of native contacts than their average sequence separation. This is in accord with the intuition that highly interdependent contacts may mutually enhance their probability of formation, thus facilitating the progress towards the native state during the folding process. This is, indeed, consistent with the negative correlation observed between cliquishness and native content, θ_m, at the transition state. It is important to stress that the presence and effects of the cooperative formation of native interactions cannot be captured by parameters based on measures of contact locality. This highlights the importance of considering all viable topologic descriptors to characterize the folding process, since they do not impact in the same way on various folding properties.

B. Optimal combined correlation

A natural question that arises is whether it is possible to combine the predicting power of cliquishness and contact order to achieve correlations with experimental folding rates and transition state placements that are better than the individual cases.

Indeed, as shown in Appendix A, it is straightforward to combine in an optimal linear way the two quantities to improve the prediction accuracy. The quantitative increment in the correlation is clearly related to the amount of independent information contained in the two topologic descriptors. Hence, an important issue is to what extent cliquishness and contact order are mutually correlated.

If, in place of a physical contact map, Δ_{ij}, one uses a random symmetric matrix, no meaningful correlation will be found. The contact maps of real proteins, however, display features that are highly non-random which reflect both (i) the physical constraints to which a compact three-dimensional chain is subject and (ii) the presence and organisation of secondary motifs[51,52,53,54].

With the aid of numeric simulations it was possible to assess the degree of interdependency of clustering coefficient of native contacts and their average sequence separation resulting from the first of the mentioned effects. This was accomplished by considering, in place of the proteins of tables I and II, 150 computer-generated compact structures respecting basic steric constraints found in real proteins (details can be found in the Methods section). As visible in the plot of Fig. 8 the level of mutual contact order-cliquishness correlation observed in these artificial structures is $r = 0.25$ which is significantly smaller than the actual correlation of the two quantities found in real proteins. In fact, the typical correlation for cliquishness and contact order (either relative or absolute) is around 0.65. Such non trivial correlation can been ascribed to the special topologic properties of naturally occurring proteins whose ramifications have been investigated in a variety of contexts[30,45,55,56,57,58]. Thus, the very presence and organization of secondary motifs in proteins makes it possible, on one hand, to exploit the native topology to predict e.g. folding rates, while on the other it limits the amount of independent information contained in different topologic descriptors.

Nevertheless, since the mutual correlation is not perfect, it is still possible to achieve, by definition, better predictions by combining cliquishness and contact order. The degree of enhancement depends also on the statistical significance of the individual starting correlations. For these reasons, the improvement is noticeable for folding rates, while it is not significant for transition state placement. For the case of two-state folders, the optimal combination yields correlations of $r = 0.86$ while for the more general case of two and three-state folding rates one has $r = 0.74$ which leads to a discernible improvement over previous cases, as visible in Fig. 7. To the best of our knowledge, this is the highest correlation recorder among similar studies involving a comparable number of entries (also including non-linear prediction schemes[27]). Due to the fact that the optimal combined correlations are found *a posteriori*, the associated values of t are no more meaningful indicators of statistical significance.

Besides the cliquishness, we have investigated other parameters that are routinely used to characterise general networks (networks of contacts in our case). In particular, we considered the "diameter" of the contact map, defined as the largest degree of separation between any two residues, and also its average value. The diameter measures the maximum number of contact that need to be traversed to connect an arbitrary pair of distinct residues. Although the contact-map diameter is an abstract object, it conveys relevant topological information about protein structure, since it measures the long-range structural organisation. We found, a posteriori, that neither the maximum, nor the average diameter, correlate in a significant manner with the folding rate or transition state placement.

III. CONCLUSIONS

We have analysed important topological descriptors of organised networks (in our case the spatial network of native contacts) that could be used, individually, or in mutual conjunction, to describe and predict experimental parameters

used to characterize the folding process. It is found that, besides the previously introduced contact order, a topologic parameter, termed cliquishness or clustering coefficient, is a powerful indicator of both the folding velocity and the transition state placement for two- and three-state folders. The predicting power of the cluquishness is that it takes into account the presence and organisation of clusters of interdependent contacts that are putatively responsible for the cooperative formation of native-like regions. This property appears well-suited to reproduce important features in the transition state that are otherwise elusive to other topologic analysis. The high statistical significance of the observed correlations testifies the strong influence of geometric structural issues on the folding process. The maximum predicting power is obtained when the topologic information contained in the cliquishness is used in combination with the contact order; this allows to reach a linear correlation as high as 0.74 with experimental folding rates recorder in 40 experimental measurements.

IV. METHODS

A. Cross correlations

The linear correlation between two sets of data, $\{x\}$ and $\{y\}$ is obtained from the normalised scalar products of the covariations:

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_j (y_j - \bar{y})^2}} \tag{4}$$

Without loss of generality, in the following we shall consider the sets of data to be with zero average and with unit norm, so that the expression of the correlation simplifies

$$r = \vec{x} \cdot \vec{y} \tag{5}$$

We now formulate the following problem. Two sets of data, $\{x\}$ and $\{y\}$ have linear correlation r_x and r_y respectively with a third (reference set), $\{z\}$. What is the maximum and minimum correlations we can expect between sets $\{x\}$ and $\{y\}$? We assume that r_x and r_y are positive since this condition can always be met by changing sign, if necessary, to the vector components.

The answer is easily found by decomposing $\{x\}$ and $\{y\}$ into their components parallel and orthogonal to $\{z\}$:

$$\vec{x} \cdot \vec{y} = b_{\|}\ c_{\|} + b_{\perp}\ c_{\perp} \tag{6}$$

Since $_{\|}\ c_{\|}$ is equal to $r_x\ r_y$, and hence is fixed, the maximum [minimum] correlation is found when $b_{\perp}$ and $c_{\perp}$ are [anti]parallel. Thus,

$$r_x\ r_y - \sqrt{(1 - r_x^2)(1 - r_y^2)} \leq r \leq r_x\ r_y + \sqrt{(1 - r_x^2)(1 - r_y^2)} \tag{7}$$

Now we turn to a different, but related problem. How can we combine linearly $\{x\}$ and $\{y\}$, so to have the maximum correlation with $\{z\}$. The generic linear combination,

$$\vec{k} = \frac{\vec{x} + b\vec{y}}{\sqrt{1 + b^2 + 2b\ \vec{x} \cdot \vec{y}}} \tag{8}$$

leads to the following correlations

$$\vec{k} \cdot \vec{z} = \frac{r_x + b\ \ r_y}{\sqrt{1 + b^2 + 2b\ \vec{x} \cdot \vec{y}}} \tag{9}$$

The maximum is achieved for

$$b = \frac{r_y - \vec{x} \cdot \vec{y}\ r_x}{r_x\ -\ \ \vec{x} \cdot \vec{y}\ \ r_y} \tag{10}$$

which yields

$$Max(\vec{k}\cdot\vec{z}) = \sqrt{\frac{r_x^2 + r_y^2 - 2\vec{x}\cdot\vec{y}\; r_x\; r_y}{1-(\vec{x}\cdot\vec{y})^2}} \tag{11}$$

B. Generation of alternative compact structures

To generate the thirty randomly-collapsed structures used in the comparison of Fig. 8, we adopted a Monte Carlo technique. The length of the artificial proteins ranged uniformly in the interval 80-110. Starting from an open conformation, each structure was modified under the action of typical MC moves (single-bead, crankshaft, pivot)[59]. A newly generated modified configuration is accepted according to the ordinary Metropolis rule. The energy scoring function is composed of two terms: The first one contains a homopolymeric part that rewards the establishment of attractive interactions (cutoff of 6.0 Å) between any pair of non-consecutive residues. The second term is introduced to penalise structure realisations with radii of gyration larger than that found in naturally-occurring proteins with the same length. The Monte Carlo evaluation is embedded in a simulated annealing scheme[60] which allows to minimise efficiently the scoring function by slowly decreasing a temperature-like control parameter.

V. ACKNOWLEDGEMENTS

We are indebted with Amos Maritan for several illuminating discussions and with Fabio Cecconi and Alessandro Flammini for a careful reading of the manuscript. Support from INFM and MURST Cofin99 is acknowledged.

[1] C. Anfinsen, Science **181**, 223 (1973).
[2] T. Creighton, *Proteins, structure and molecular properties* (W.H.Freeman and Company, New York, 1993), 2nd ed.
[3] C. Branden and J. Tooze, *Introduction to protein structure* (Garland Publishing, New York, 1991).
[4] S. E. Jackson and A. R. Fersht, Biochemistry **30**, 10428 (1991).
[5] S. E. Jackson, Folding and Design **3**, R81 (1998).
[6] G. S. Huang and T. G. Oas, Biochemistry **34**, 3884 (1995).
[7] R. E. Burton, G. S. Huang, M. A. Daugherty, P. W. Fullbright, and T. G. Oas, J. Mol. Biol. **163**, 311 (1996).
[8] B. B. Kragelund, C. V. Robinson, J. Knudsen, C. M. Dobson, and F. M. Poulsen, Biochemistry **34**, 7217 (1995).
[9] B. B. Kragelund, P. Hojrup, M. S. Jensen, Schjerling, E. Juul, J. Knudsen, and F. M. P. FM, J. Mol. Biol. **256**, 187 (1996).
[10] G. A. Mines, T. Pascher, S. C. Lee, J. R. Winkler, and H. B. Gray, Chem. Biol. **3**, 491 (1996).
[11] N. Schonbrunner, K. P. Kofler, and T. Kiefhaber, J. Mol. Biol. **268**, 526 (1997).
[12] K. L. Reid, H. M. Rodriguez, B. J. Hillier, and L. M. Gregoret, Protein Sci. **7**, 470 (1998).
[13] V. P. Grantcharova and D. Baker, Biochemistry **36**, 15685 (1997).
[14] V. Villegas, A. Azuaga, L. Catasus, D. Reverter, P. L. Mateo, F. X. Aviles, and L. Serrano, Biochemistry **34**, 15105 (1995).
[15] S. Khorasanizadeh, I. D. Peters, T. R. Butt, and H. Roder, Nat. Struct. Biol. **3**, 193 (1993).
[16] M. L. Scalley, Q. Yi, H. D. Gu, A. McCormack, J. R. Yates, and D. Baker, Biochemistry **36**, 3373 (1997).
[17] M. Silow and M. Oliveberg, Biochemistry **36**, 7633 (1997).
[18] N. A. J. Van Nuland, W. Meijberg, J. Warner, V. Forge, R. M. Schee, G. T. Robillard, and C. M. Dobson, Biochemistry **37**, 622 (1998).
[19] N. Taddei, F. Chiti, P. Paoli, T. Fiaschi, M. Bucciantini, M. Stefani, C. M. Dobson, and G. Ramponi, Biochemistry **38**, 2135 (1999).
[20] N. Ferguson, A. P. Capaldi, R. James, C. Kleanthous, and S. E. Radford, J. Mol. Biol. **286**, 1597 (1999).
[21] C. K. Smith, Z. M. Bu, J. M. S. K. S. Anderson, D. M. Engelman, and L. Regan, Protein Sci **5**, 2009 (1996).
[22] B. Kuhlman, D. L. Luisi, P. A. Evans, and D. P. Raleigh, J. Mol. Biol. **284**, 1661 (1998).
[23] S. J. Hamill, A. E. Meekhof, and J. Clarke, Biochemistry **37**, 8071 (1998).
[24] Y.-J. Tan, M. Oliveberg, and A. R. Fersht, J. Mol. Biol. **264**, 377 (1996).
[25] K. W. Plaxco, K. T. Simons, and D. Baker, J. Mol. Biol. **277**, 985 (1998).
[26] D. A. Debe and W. A. Goddard III, J. Mol. Biol. **294**, 619 (1999).
[27] A. R. Dinner and M. Karplus, Nat. Struct. Biol. **8**, 21 (2001).
[28] D. N. Ivankov and A. Finkelstein, Biochemistry **40**, 9957 (2001).
[29] R. Aurora, T. P. Creamer, R. Srinivasan, and G. D. Rose, J. Mol. Biol. **272**, 1413 (1997).
[30] A. Maritan, C. Micheletti, and J. R. Banavar, Phys. Rev. Lett. **84**, 3009 (2000).
[31] A. P. Capaldi and S. E. Radford, Curr. Op. Str. Biol. **8**, 86 (1998).
[32] J. Clarke, E. Cota, S. B. Fowler, and S. J. Hamill, Structure **7**, 1145 (1999).

[33] B. Bollabas, *Random Graphs* (Academic, London, 1985).
[34] D. J. Watts and S. H. Strogatz, Nature **393**, 440 (1998).
[35] S. H. Strogatz, Nature **410**, 268 (2001).
[36] N. Go and H. A. Scheraga, Macromolecules **9**, 535 (1976).
[37] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, Proteins **21**, 167 (1995).
[38] P. G. Wolynes, J. N. Onuchic, and D. Thirumalai, Science **267**, 1619 (1995).
[39] K. A. Dill and H. S. Chan, Nature Structural Biology **4**, 10 (1997).
[40] Dobson, C. M., Sali, A., Karplus, and M, Angew. Chem. Int. Edit. **37**, 868 (1998).
[41] A. Sali, E. Shakhnovich, and M. Karplus, Nature **369**, 248 (1994).
[42] H. S. Chan and K. A. Dill, J. Chem. Phys. **92**, 3118 (1990).
[43] F. Cecconi, C. Micheletti, P. Carloni, and A. Maritan, Proteins: Structure Function and Genetics **43**, 365 (2001).
[44] D. A. Baker, Nature **405**, 39 (2000).
[45] C. Micheletti, J. R. Banavar, A. Maritan, and F. Seno, Phys. Rev. Lett. **82**, 3372 (1999).
[46] C. Clementi, H. Nymeyer, and J. N. Onuchic, J. Mol. Biol. **298**, 937 (2000).
[47] O. V. Galzitskaya and A. V. Finkelstein, Proc. Natl. Acad. Sci. USA **96**, 11299 (1999).
[48] V. Munoz, E. R. Henry, J. Hofrichter, and W. A. Eaton, Proc. Natl. Acad. Sci. USA **95**, 5872 (1999).
[49] E. Alm and D. Baker, Proc. Natl. Acad. Sci. USA **96**, 11305 (1999).
[50] D. K. Klimov and D. Thirumalai, Proc. Natnl. Acad. Sci. USA **97**, 7254 (2000).
[51] C. Chothia, Annu. Rev. Biochem. **53**, 537 (1984).
[52] C. Chothia, Nature **357**, 543 (1992).
[53] M. Levitt and C. Chothia, Nature **261**, 552 (1976).
[54] G. D. Rose and J. P. Seltzer, J. Mol. Biol. **113**, 153 (1977).
[55] A. Maritan, C. Micheletti, A. Trovato, and J. R. Banavar, Nature **406**, 287 (2000).
[56] M. Denton and C. Marshall, Nature **410**, 417 (2001).
[57] N. G. Hunt, L. M. Gregoret, and F. E. Cohen, J. Mol. Biol. **241**, 214 (1994).
[58] N. D. Socci, W. S. Bialek, , and J. N. Onuchic, Phys. Rev. E **49**, 3440 (1994).
[59] A. D. Sokal, Nuclear Physics **B47**, 172 (1996).
[60] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, Science **220**, 671 (1983).
[61] A. Viguera, J. Martinez, V. Filimonov, P. Mateo, and L. Serrano, Biochemistry **33**, 2142 (1994).
[62] C. Chan, Y. Hu, S. Takahashi, D. L. Rousseau, W. A. Eaton, and J. Hofrichter, Proc. Natl. Acad. Sci. USA **94**, 1779 (1997).
[63] K. Plaxco, J. Guijarro, C. Morton, M. Pitkeathly, I. Campbell, and C. Dobson, Biochemistry **37**, 2529 (1998).
[64] D. Perl, C. Welker, T. Schindler, K. Schroeder, M. A. Marahiel, R. Jaenicke, and F. Schmid, Nat. Struct. Biol. **5**, 229 (1998).
[65] I. Guijarro, C. Morton, K. W. Plaxco, I. D. Campbell, and C. Dobson, J. Mol. Biol. **276**, 657 (1998).
[66] J. Clarke, S. J. Hamill, and C. M. Johnson, J. Mol. Biol. **270**, 771 (1997).

Protein	Length	Family	ln K_f	θ_m	Cliquishness ($d = 4.6$ Å)
1shg[61]	57	α	2.10	0.69	0.546
1lmb[6,7]	80	α	8.50	0.46	0.555
2abd[8,9]	86	α	6.55	0.61	0.550
1imq[26]	86	α	7.31	—	0.545
1ycc[10]	103	α	9.61	0.34	0.548
1hrc[62]	104	α	7.94	0.47	0.538
1hrc, horse, oxidized Fe^{III}[62]	104	α	5.99	0.40	0.538
2gb1[21]	56	α/β	6.26	—	0.553
2ptl[16]	61	α/β	4.22	0.75	0.551
2ci2[4]	65	α/β	3.87	0.61	0.535
1cis[24]	66	α/β	3.87	0.61	0.540
1hdn[18]	85	α/β	2.70	0.64	0.526
1aye[14]	92	α/β	6.80	0.74	0.554
1urn[17]	96	α/β	5.73	0.55	0.520
1aps[19]	98	α/β	-1.47	0.79	0.511
1fkb[5]	107	α/β	1.46	0.67	0.512
2vik[5]	126	α/β	6.80	0.73	0.546
1srl[13]	56	β	4.04	0.69	0.536
1shf.a[63]	59	β	4.55	0.68	0.534
1tud[26]	60	β	3.45	—	0.531
1csp[64]	67	β	6.04	0.85	0.539
1mjc[12]	69	β	5.23	0.91	0.532
3mef[12]	69	β	5.30	0.94	0.531
2ait[11]	74	β	4.20	0.65	0.542
1pks[65]	76	β	-1.05	0.60	0.545
1ten[66]	89	β	-1.10	0.76	0.511
1fnf, 9FN-III[32]	90	β	-0.90	0.63	0.513
1wit[32]	93	β	0.41	0.70	0.526
1fnf, 10FN-III[32]	94	β	5.00	0.65	0.536

TABLE I: List of proteins known to fold via a two-state mechanism. The experimental quantities K_F (s^{-1}) and θ_m are desumed from the cited references. The reported cliquishness values are calculated for the cutoff $d = 4.6$ Åyielding optimal correlations against folding rates.

Protein	Length	Family	ln K_f	θ_m	Cliquishness ($d = 4.6$ Å)
1bta	89	α	3.40	0.87	0.532
1ubq	76	α/β	5.90	0.59	0.532
1bni	108	α/β	2.60	0.88	0.524
1hel	129	α/β	1.30	0.75	0.507
3chy	128	α/β	1.0	0.71	0.512
1dk7	146	α/β	0.80	0.78	0.513
2rn2, Urea, pH 5.5	155	α/β	-0.50	0.80	0.502
2rn2, GdnHCl, pH 5.5	155	α/β	1.40	0.63	0.502
1php.n	175	α/β	2.30	0.84	0.505
1hng, pH 7.0	97	β	1.80	0.68	0.502
1hng, pH 4.5	97	β	2.63	0.62	0.502

TABLE II: List of proteins known to fold via a three-state mechanism. The experimental quantities K_F (s^{-1}) and θ_m are desumed from[5]. The reported cliquishness values are calculated for the cutoff yielding optimal correlations against folding rates.

Figure captions

- Fig. 1. Correlation of cliquishness, relative and absolute contact order against folding rates of two-state folders. The values of the correlation coefficients are plotted as a function of the cutoff, d, used in the definition of the contact map.
- Fig. 2. Correlation of cliquishness, relative and absolute contact order against transition state placement of two-state folders. The values of the correlation coefficients are plotted as a function of the interaction cutoff, d.
- Fig. 3. Correlation of cliquishness, relative and absolute contact order against folding rates of two- and three-state folders. The values of the correlation coefficients are plotted as a function of the cutoff, d, used in the definition of the contact map.
- Fig. 4. Correlation of cliquishness, relative and absolute contact order against transition state placement of two- and three-state folders. The values of the correlation coefficients are plotted as a function of the interaction cutoff, d.
- Fig. 5. Scatter plot of cliquishness (left) and absolute contact order (right) versus folding rates of the 40 entries of Tables 1 and 2. The used values of d are the optimal ones reported in the text. Filled circles, open squares and starred points denote proteins belonging to the α, α/β and β families, respectively.
- Fig. 6. Scatter plot of cliquishness (left) and absolute contact order (right) versus θ_m of the entries of Tables 1 and 2. The used values of d are the optimal ones reported in the text. Filled circles, open squares and starred points denote proteins belonging to the α, α/β and β families, respectively.
- Fig. 7. Scatter plot of the logarithm of folding rates for the entries of Tables 1 and 2, against data from optimally combined cliquishness and contact order. The optimal linear superposition, see Methods, is obtained for $b = 0.7$, ($\{x\}$ and $\{y\}$ being the cliquishness and contact order data respectively. Filled circles, open squares and starred points denote proteins belonging to the α, α/β and β families, respectively.
- Fig. 8. Scatter plot of average cliquishness versus absolute contact order, for randomly collapsed structures generated by stochastic numerical methods.

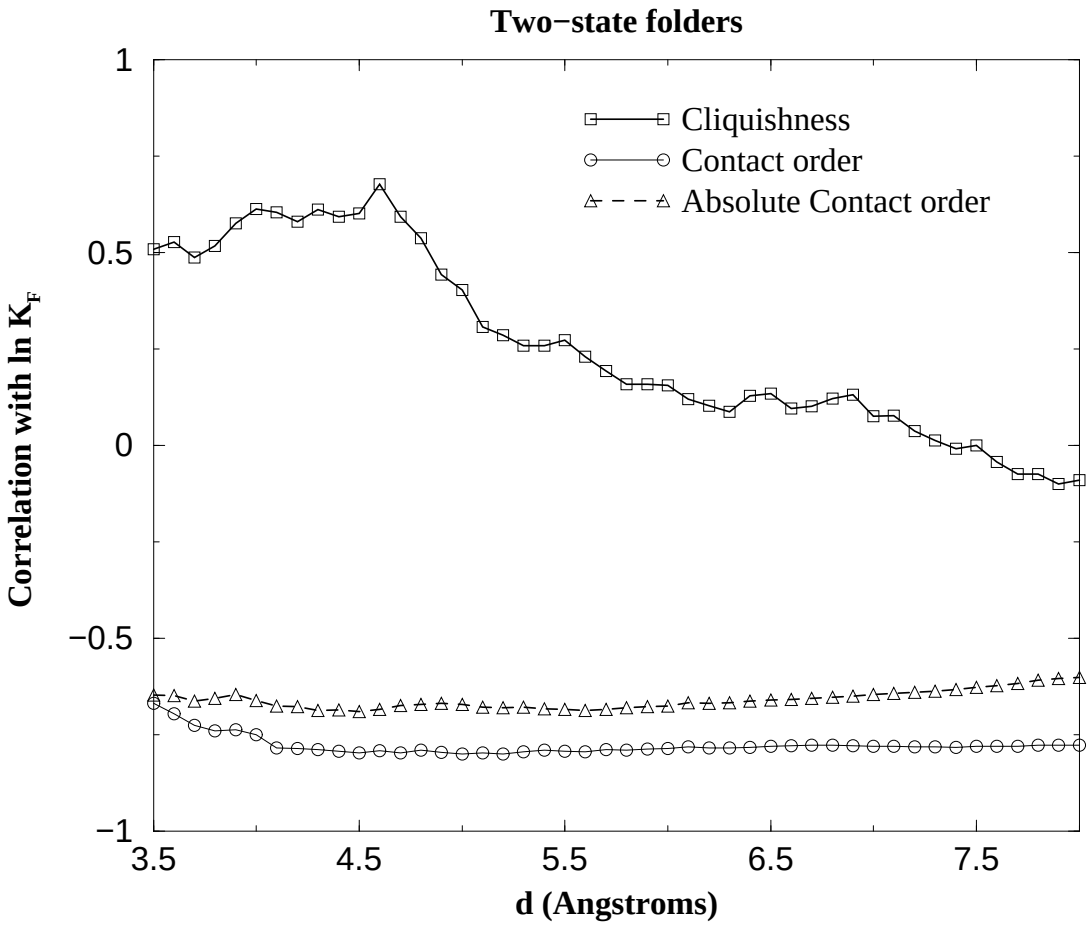
Two−state folders
Cliquishness
Contact order
Absolute Contact order
Correlation with ln K_F
d (Angstroms)
1
0.5
0
−0.5
−1
3.5
4.5
5.5
6.5
7.5

FIG. 1:

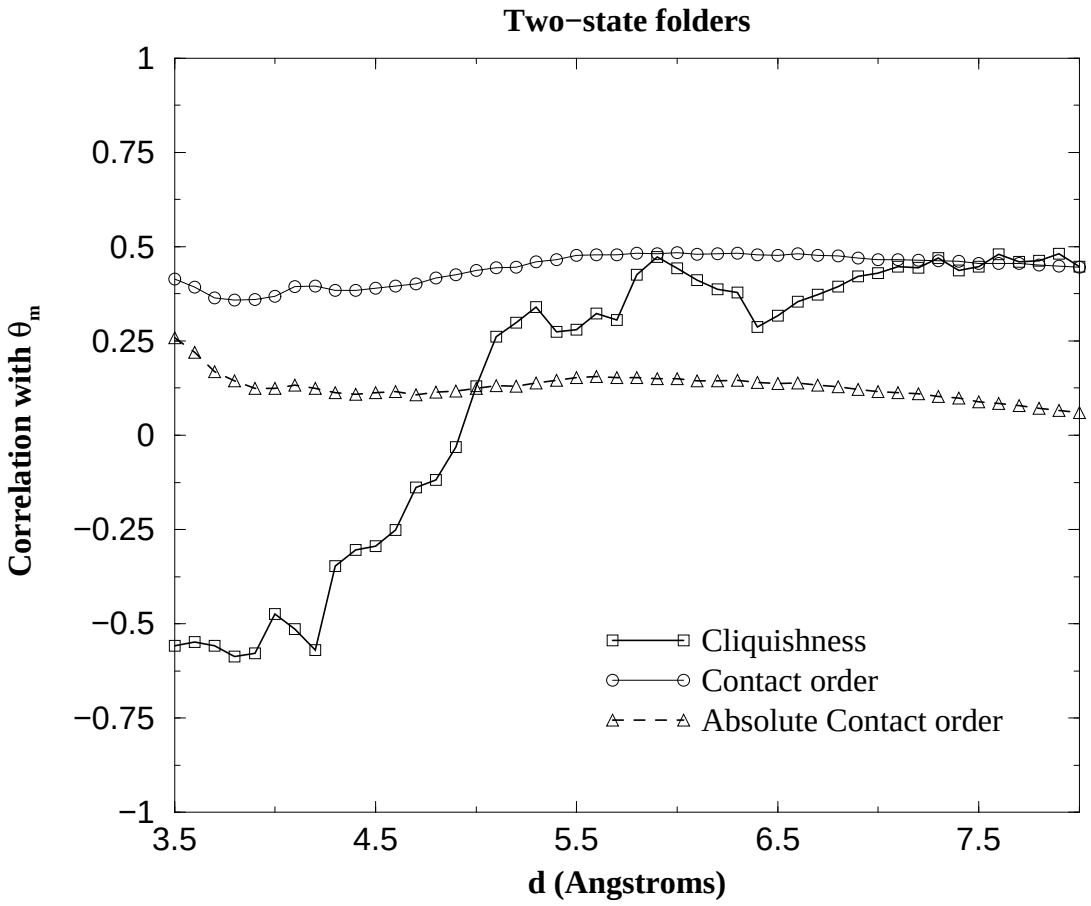
Two−state folders
Correlation with θm
d (Angstroms)
Cliquishness
Contact order
Absolute Contact order
1
0.75
0.5
0.25
0
−0.25
−0.5
−0.75
−1
3.5
4.5
5.5
6.5
7.5

FIG. 2:

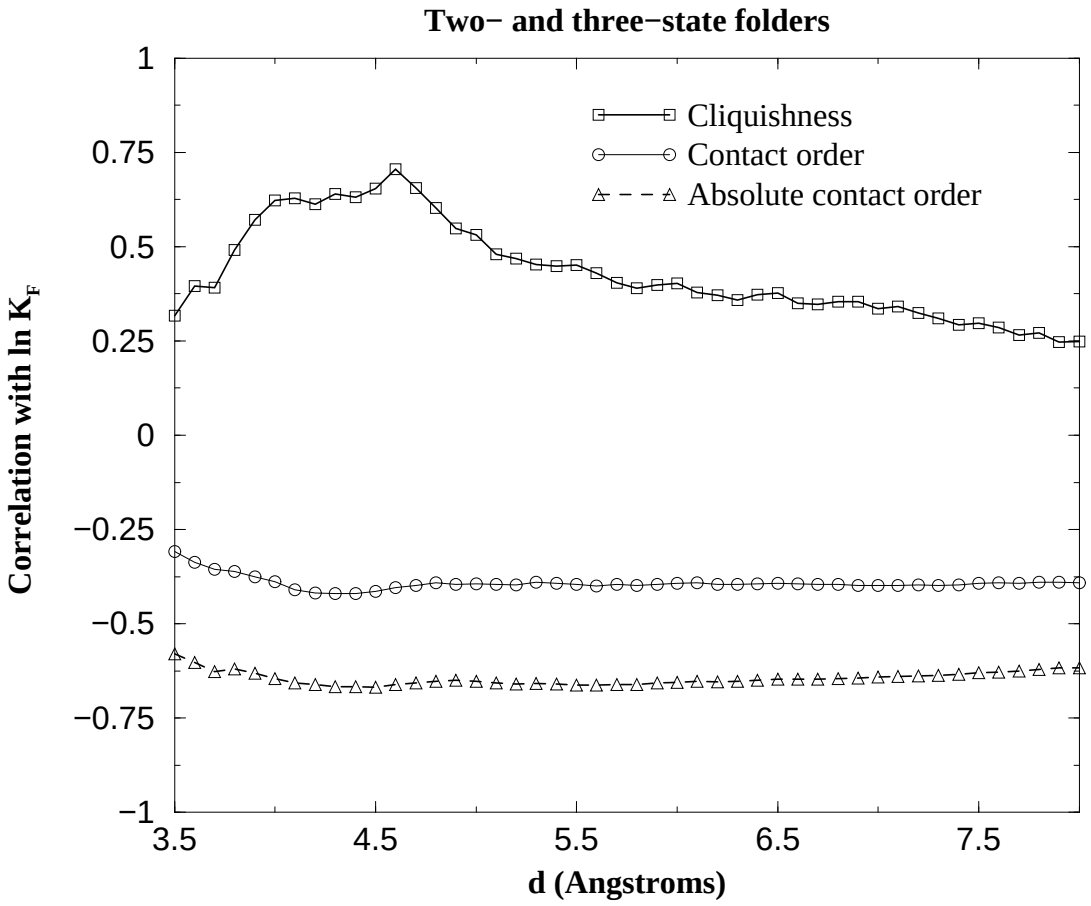
Two− and three−state folders
Cliquishness
Contact order
Absolute contact order
Correlation with ln K_F
d (Angstroms)
1
0.75
0.5
0.25
0
−0.25
−0.5
−0.75
−1
3.5
4.5
5.5
6.5
7.5

FIG. 3:

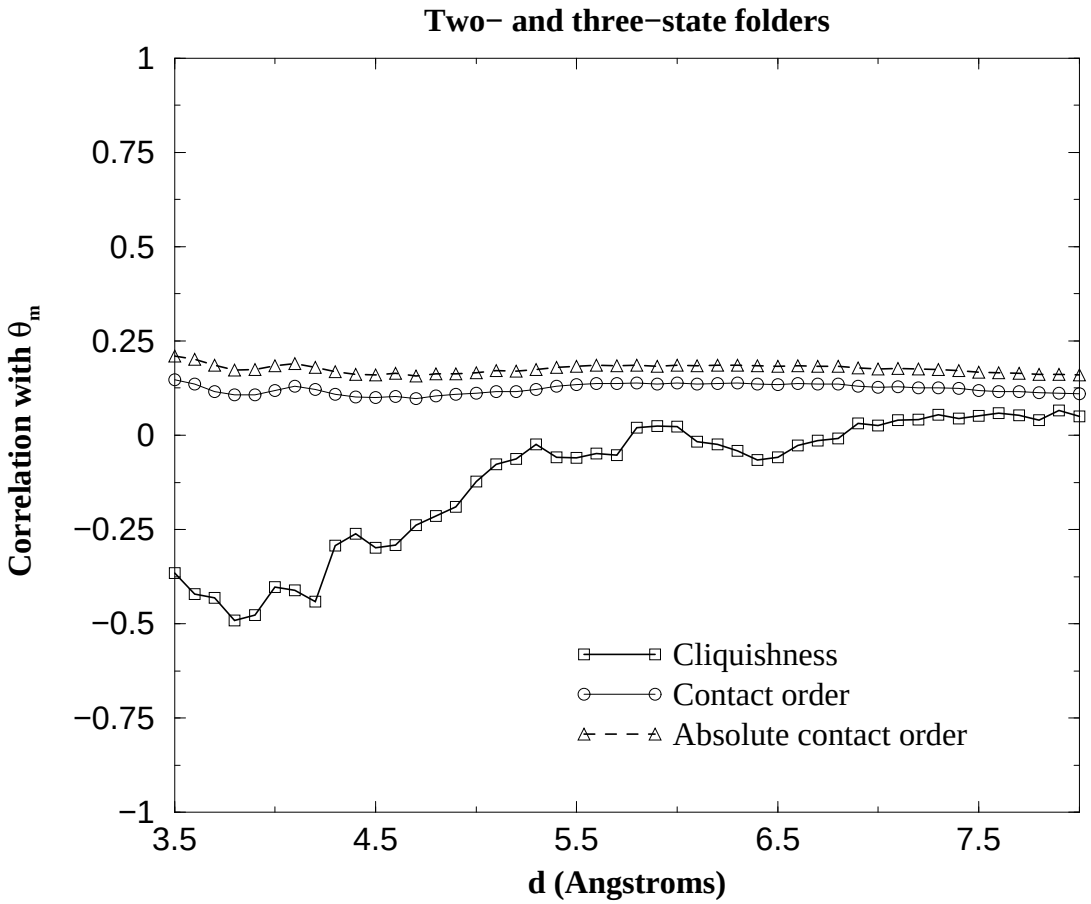
Two− and three−state folders
Correlation with θm
d (Angstroms)
Cliquishness
Contact order
Absolute contact order

FIG. 4:

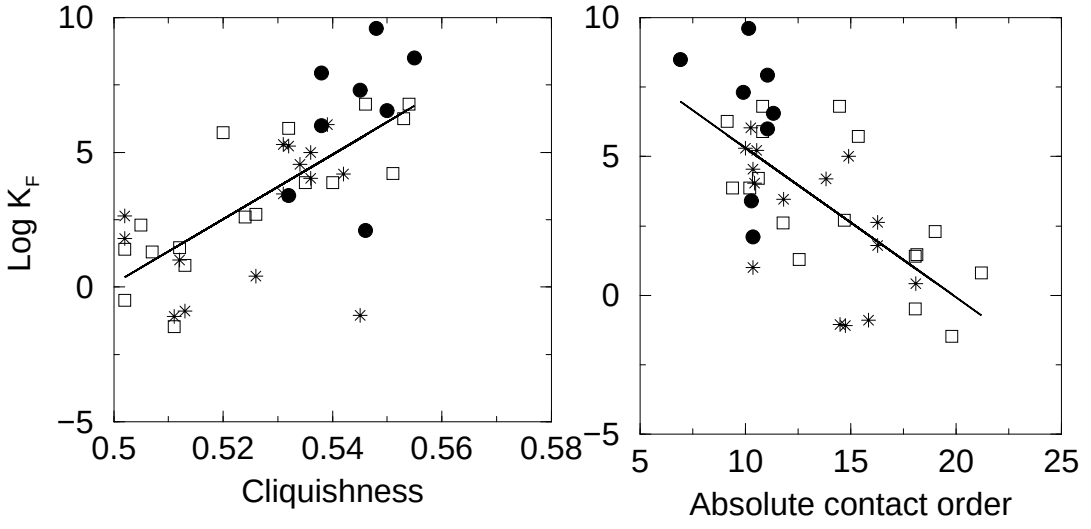
Log KF
Cliquishness
Absolute contact order

FIG. 5:

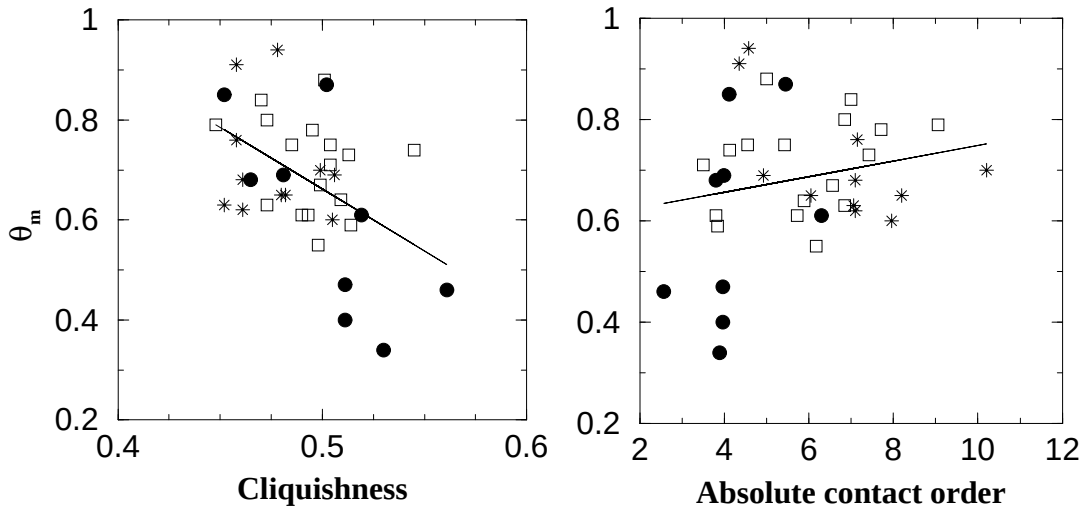
1
0.8
0.6
0.4
0.2
θm
0.4
0.5
0.6
Cliquishness
1
0.8
0.6
0.4
0.2
2
4
6
8
10
12
Absolute contact order

FIG. 6:

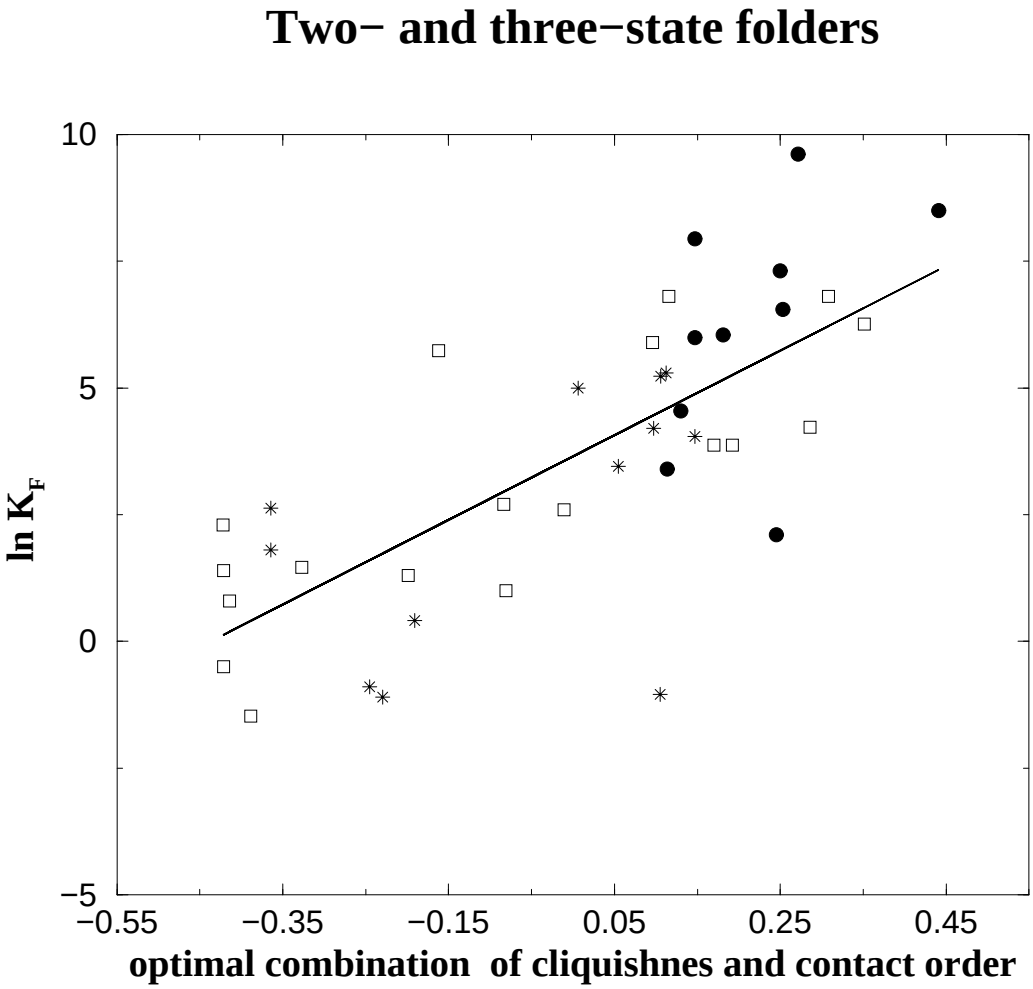
Two− and three−state folders
10
5
0
−5
ln KF
−0.55
−0.35
−0.15
0.05
0.25
0.45
optimal combination of cliquishnes and contact order

FIG. 7:

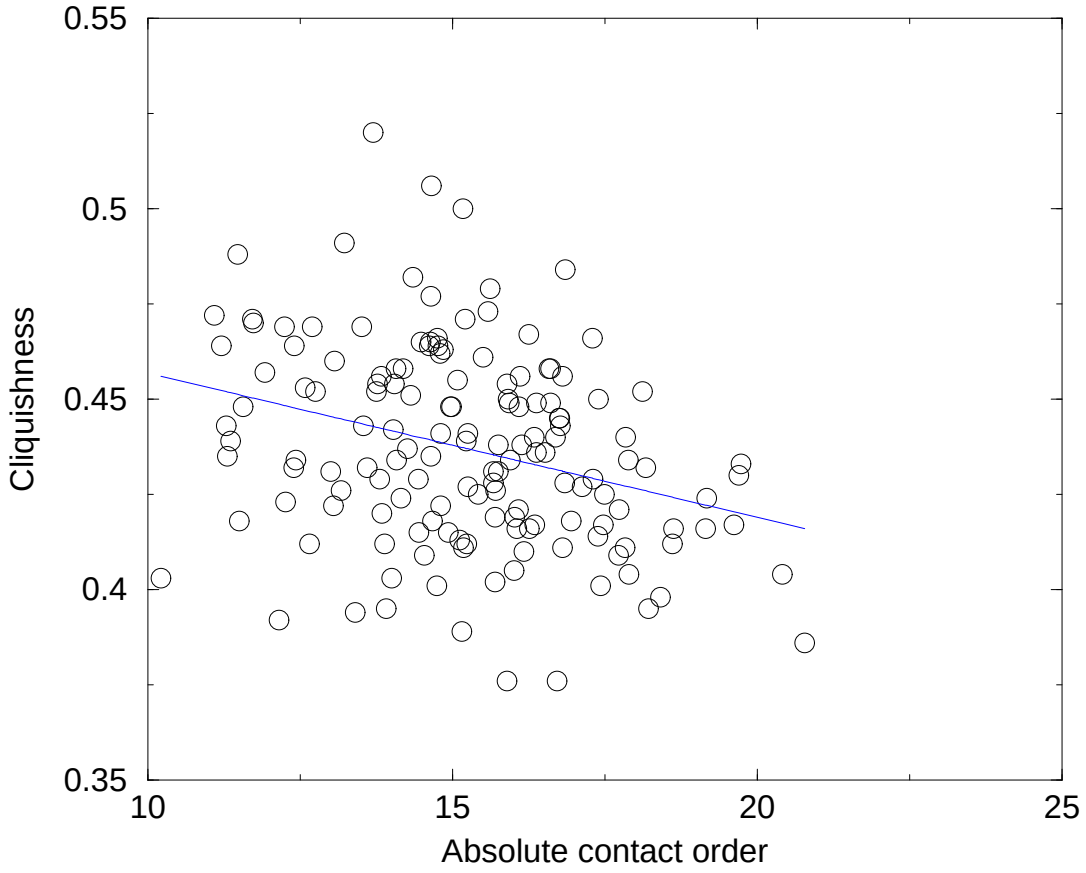
0.55
0.5
0.45
0.4
0.35
Cliquishness
10
15
20
25
Absolute contact order

FIG. 8: